Honours Meeting

July 20th, 2020

The general answer is:

Conf. var. introduce biases into the datasets:

* Normalize out so they contribute with minimal information distributions
* Equally biasing all of the data
* gets challenging depending on the feature
* ORE DATA
  + Order of magnitude difference in size
  + Volume impacts features
  + Get rid of it by
  + Plot two data pieces against each other
  + Lag-lag curves \*\*\*\* Not super descriptive
  + Correlation – linear relationship
* Training set, anything u think is key feature you don’t want it keying on you want it to have a uniform distribution
* Tank example – sunny days v.s cloudy days
* Correcting for unbalanced dataset may increase error rate
* NN/deep learning can lock on to specifics of the data
  + These systems in practice become difficult
* Hoping to be able to do the classification is vastly reduced space (or you don’t have enough data)

New feature space definition:

* Do it both ways
  + Bin low medium high AQ areas
  + Not merging that data
* Median air quality more important than surges
  + Can look for trends here
  + Each way you look at the data exposes different issues and outcomes
  + Takes a lot of hands on experience to know what approach to take
* Where they are in terms of percentile in terms of one another
* Once AQ past another threshold you’re screwed? Wouldn’t be surprised if that was the case
* Expected death rates – how many died above this. Something else surging death……

Python libs have similar support for vector & matricies